DOCUMENT RESUME

ED 410 885 HE 030 482

AUTHOR Sadler, William E.; Cohen, Frederic L.; Kockesen, Levent
TITLE Factors Affecting Retention Behavior: A Model To Predict At-Risk Students. AIR 1997 Annual Forum Paper.
PUB DATE 1997-05-00
NOTE 22p.; Paper presented at the Annual Forum of the Association for Institutional Research (37th, Orlando, FL, May 18-21, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Persistence; *College Freshmen; Dropouts; *High Risk Students; Higher Education; Identification; Institutional Research; Models; Predictive Measurement; Predictive Validity; Predictor Variables; Regression (Statistics); Research Methodology; School Holding Power
IDENTIFIERS *AIR Forum; Logistic Regression; *New York University

ABSTRACT

This paper describes a methodology used in an on-going retention study at New York University (NYU) to identify a series of easily measured factors affecting student departure decisions. Three logistic regression models for predicting student retention were developed, each containing data available at three distinct times during the first semester: first, prior to the start of the fall semester; second, after the third week of classes; and third, at the end of the first semester. A method of identifying appropriate variables for inclusion in the logistic regression model is discussed as well as a rationale for choosing different cut points to classify the logit results. The study followed Fall 1994 and Fall 1995 freshmen (n=2209) among whom 272 students did not return to NYU a year after entry. Variables were grouped into six general categories describing: (1) family background/individual attributes; (2) pre-college schooling; (3) institution commitment; (4) first-term academic integration; (5) first-term social integration; and (6) first-year finances. The study found all three models were reasonably effective in identifying high risk students using various probability cutoff points. It concluded that use of all three models to identify students at risk at the three different times would allow for an optimum intervention strategy. (BF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 410 885

Factors Affecting Retention Behavior:

A Model to Predict At-Risk Students[1]

by

William E. Sadler,
Graduate Assistant
Office for Enrollment Research and Analysis
New York University
7 East 12th Street, Suite 615
New York, NY 10003
(212) 998-4420

Fredric L. Cohen,
Director
Office for Enrollment Research and Analysis
New York University
7 East 12th Street, Suite 615
New York, NY 10003
(212) 998-4415

Levent Kockesen,
Teaching Assistant
Economics Department
New York University
269 Mercer Street, Suite 700
New York, NY 10003
(212) 998-4420

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
AIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

[1] Presented to the Association for Institutional Research Annual Forum, Orlando, Florida, May 1997. For questions or comments, please contact William E. Sadler at New York University, 7 East 12th Street, Suite 615, New York, NY 10003 or by e-mail at JAWES@UCCVM.NYU.EDU.

IR030482

BEST COPY AVAILABLE

for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-Seventh Annual Forum of the Association for Institutional Research held in Orlando, Florida, May 18-21, 1997. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Jean Endo
Editor
AIR Forum Publications

Factors Affecting Retention Behavior: A Model to Predict At-Risk Students

Abstract

Institutional researchers routinely collect and report numerical data on student retention, but in so doing rarely scratch the surface when addressing the problem of student attrition. This paper describes the results of an on-going retention study at New York University to identify a series of easily measured factors affecting student departure decisions. Three logistic regression models were developed, each containing data available at three distinct times during the first semester, to predict freshmen at risk for dropping out. A method for identifying appropriate variables for inclusion in the logit model is discussed as well as a rationale for choosing different cut points to classify the logit results.

Each semester, institutional researchers are asked by members of their institution's senior administration to "report" on retention statistics. All too often, these researchers fail to really scratch the surface when addressing the problem of student attrition; they prepare charts and graphs each semester, provide descriptive statistics, make a presentation or two, and move on to the next project. Unfortunately, by not delving deeper into the question of attrition, institutional decision makers are forced to rely on anecdotal evidence about why students fail to retain. When this happens, programs designed to prevent students from dropping out may be inappropriately designed.

What alternatives are there? If one could provide a predictive model to institutional decision makers, then appropriate interventions could be created, targeted directly towards those most at risk for leaving. What factors might one include in such a model? Certainly there are many data elements which could be useful in determining students' likelihood to attrit; some available in the pre-enrollment time frame and some available only post-enrollment. Which data elements are included in any model is really a campus specific issue, determined both by the availability of the data for a substantial part of the student population, as well as when campus leaders might want the model to be run.

This paper will focus on the process which was followed at New York University in the development of a model for student attrition from one of its undergraduate colleges. It is meant to be an example of the steps any institution might follow in developing an early warning system, and so we will review the decisions and choices made along the way by NYU. This model, as presented here, is ***not*** meant to be prescriptive for any other institution; rather, we hope to provide the tools and background for others to provide the same analysis on their own campuses.

Prior Research on Student Retention

A significant body of literature exists on the issue of undergraduate student retention; three major researchers are highlighted here. The first significant research on the issue of student retention was by Spady (1970). It provided the first theoretical model of the dropout process in higher education. This model proposes that social integration (manifested by shared group values, academic performance, normative congruence and support of friends) increases institutional commitment, which, in turn, reduces the likelihood that a student will attrit. The model suggests that student background characteristics (family and personal characteristics and skills) also combine to influence the attrition process.

Building on Spady's work, Tinto (1975) provided a definitive theoretical model that described the process of student integration into academic and social systems at a particular institution. It encompasses: pre-college attributes; student goals and commitments prior to college entry; formal and informal college experiences; personal/normative integration; and, goals and commitments after college entry--all culminating in a student's decision to stay or depart.

Cabrera, Nora, and Casañeda (1992) brought the issue of finances to the attrition literature. They found that while finances do not have a direct effect on a student's persistence, they do have indirect effects on persistence through intervening variables, more specifically, through a "student's academic integration, socialization processes, as well as his or her resolve to persist in college" (p. 589).

Overview of the Study

In Fall 1995, the Office for Enrollment Research and Analysis at New York University began to examine what was behind one particular college's raw retention numbers when its Dean asked that we undertake a project to identify factors that affected the retention behavior

of his college's students. This project would help to bring the retention numbers to a more personal level so that the institution could better respond to the needs of its students.

The first goal of this project was to develop a model, which could be used early in the freshman year, to predict which students are at-risk for attriting from the college. This group of at-risk students would then be contacted by the staff of the college so that the needs of individual students could be identified and addressed in a one-on-one basis. The ultimate goal is for the college to be able to identify and interact with at-risk students as early as possible, thereby reducing the likelihood that they will leave the institution. This paper will discuss the steps taken in pursuit of the first project goal, developing the model to predict at-risk students.

Preliminary Model and Study Design

The goal to provide a predictive model for at-risk students that could be utilized early in the freshmen year placed some limits on the data that would be available for analysis. Data collection needed to be centralized and simple, with few demands on staff and students alike. To this end, it was determined to utilize data that could be retrieved from the college's databases as well as data from the University's Student Information System, where information from the files of the Undergraduate Admissions Office, Financial Aid Office, Registrar's Office and Bursar's Office are maintained. Limitations on the availability of data from the college prevented us from examining cohorts prior to the class which entered in the Fall of 1994. Therefore, our sample included data for the Fall 1994 and Fall 1995 entering freshman cohorts (N=2209). From this group, 272 students did not return to NYU one year after entry, i.e. in the Fall of the sophomore year.

Variables and Their Indicators

The theoretical model of attrition used in this study focuses on the role of pre-entry attributes and institutional experiences from high school graduation through the first semester of college. The variables are grouped into six general categories describing: family background/individual attributes, pre-college schooling, institutional commitment, first-term academic integration, first-term social integration and first-year finances (see Table 1).

Methods

A logistic regression model to predict second fall semester enrollment was developed consistent with the method suggested by Hosmer and Lemeshow (1989). Logistic regression was chosen because it allows for easier model building when the dependent variable is dichotomous (yes/no, 1/0), as it is in this case (retained or attrited). We recognize that our dependent variable, retention to the second fall, is defined simply. For example, we include in the "out of attendance" category students who are on a leave of absence, as well as students who are ultimately stop-outs, and not drop-outs (i.e. they return in a later semester). However, research at NYU has shown that only about one-fourth of students in the college who take an official leave ultimately return. Similarly, we have found that less than 6% of stop outs return to the University. Based on this data, we can accept the simpler definition of attrition, which treats each of these (leave of absence, stop-out and drop-out) as being the same.

In an effort to identify independent variables appropriate for inclusion into the logistic regression model, a bivariate analysis was conducted for each potentially important variable. This was accomplished through the use of contingency tables for dichotomous variables where the likelihood ratio chi-square test was employed to determine the level of association between the independent variables and second-fall enrollment status (dependent variable). For

Table 1

Description of Variables

Category	*Variable*	*Description*[a]
Family Background/ Individual Attributes	SEX	Female?
	AGE	Age as of August 31 of entrance term
	WHITE	White?
	BLACK	Black?
	HISPANIC	Hispanic?
	ASIAN	Asian?
	FOREIGN	Foreign student?
	F1_PAGI	First year parent's adjusted gross income
	NYC	ZIP code when admitted is in one of the 5 boroughs of NYC?
	NYAREA	ZIP code when admitted is in one of 17 counties surrounding NYC?
Pre-College Schooling	F1_RANK	First year financial aid rank
	HS_GPA	High school GPA
	HS_PRANK	High school percentage rank in class
	SAT_COMB	SAT total score
	SAT_MATH	SAT math score
	SAT_VERB	SAT verbal score
	T1_CTRHR	Transfer hours posted at start of first term (includes AP credit)
Institutional Commitment	SUMORIE	Attended Summer orientation (versus Fall orientation)?
	EARLYDEC	Did student apply for an early decision admissions decision?
Academic Integration (all first term)	T1_CATHR	Cumulative attempted hours
	T1_ATTHR	Term attempted hours
	T1_EARHR	Term earned hours
	T1_TEGPA	Term GPA
	T1_UEHR	Term unearned hours (T1_ATTHR minus T1_EARHR)
	T1_MTEA	Mid-term grades--number of courses with "excessive absences"
	T1_MTSA	Mid-term grades--number of courses with "satisfactory progress"
	T1_MTUN	Mid-term grades--number of courses with "unsatisfactory progress"
	T1_MTUE	Mid-term grades--number of courses with "not able to evaluate"
	UNDC	Undecided major?
Social Integration (all first term)	T1_RESID	Live in on-campus residence hall during first term?
	FEVENT	Number of freshman-targeted programs student attended first term
	FDEAN	Did student meet with freshman dean during first term?
	T1_FTPT	Was student full-time during first term?
Finances (all first year)	F1_UNMET	Amount of unmet financial need in first year
	T1_GRANT	Did student receive institutional (non-portable) grant aid?
	T1_GRAMT	Amount of institutional grant aid
	T1_LOAN	Did student receive institutional (non-portable) loans?
	T1_LOAMT	Amount of institutional loans
	T1_TUITR	Did student receive tuition remission benefits?
	T1_BAL	Student's first term bursar balance at end of third week of classes
	F1_NBRAT	Financial aid received as percentage of student's cost of attendance

[a] Those variables whose description ends with a question mark are dummy variables which take the value of "1" if the answer is "Yes" and "0" otherwise.

BEST COPY AVAILABLE

continuous independent variables, independent sample t-tests were conducted to compare the two groups of students (those who attrited by their second fall semester versus those who retained).

Variables that showed predictive potential (by having a p-value < 0.25) based on these analyses were then entered into a logistic regression analysis. Two techniques for building the logistic regression equation were used: forward selection with a test for backward elimination as well as backward elimination with a test for subsequent forward selection of variables. In both cases an alpha level for entry into the equation of p_E= 0.20 and an alpha level for removal from the equation of p_R= 0.25 were utilized. Following the development of a preliminary logistic regression equation, interaction effects among the variables were tested and subsequently included into the equation based on the same criteria noted above.

Results

Based on the bivariate analysis, every variable included in Table 1 showed predictive potential with the *exception* of the following: number of courses where professors were unable to evaluate performance for mid-term advisory grades; total SAT score; parent's adjusted gross income; whether a student was Black, Hispanic , foreign or full-time; whether the student received institutional (non-portable) loans; and the amount of institutional loan aid that a student received. These variables, which showed no predictive potential, were excluded from any further analysis.

We then attempted to build logistic regression models that predicted which students would retain, based upon variables that would be available to the institution at four distinct times: (1) prior to the start of the fall semester; (2) after the fall semester "census date" (end of third week of classes); (3) after mid-term advisory grades are given by the faculty; and (4) at the

end of the first semester. We successfully achieved predictive models for three out of the four stages (see Table 2), the exception being a model that utilized mid-term advisory grades.

The first model, representing variables that would be available prior to the start of the fall semester, showed that the following factors increased the odds of retention: receiving tuition remission benefits; being from New York City; being of Asian descent; having a higher high school grade point average; attending orientation in the summer; being younger; and not being undecided about an undergraduate major. The amount of unmet financial need a student had as well as the interaction between unmet need and being a New York City resident increased the overall fit of the logit model but did not directly increase or decrease the odds that the student would be retained into the second year.

The second model, representing variables that would be available at the end of the third week of classes, showed that all of the factors that had positive influences on retention in the first model continued to be positive influences here. Other items that had positive influences on retention included: having a higher percentage of the student's financial need met by financial aid; attempting a larger number of credit hours during the first semester; and having a higher number of transfer or advanced placement credits.

Although not having a positive or negative influence on retention, four additional items were included in the model to improve the overall fit, including two interaction variables. They were: the amount of institutional grant aid the student received; the amount of the student's bursar balance at the end of the third week of classes; the interaction between the number of transfer/advanced placement credit hours and the amount of institutional grant aid received; as well as the interaction between the amount of institutional grant aid received and the percentage of financial need met by the student's financial aid package. These four variables

Table 2

Logistic Regression Results--Odds Ratios (e^{β}) of Retention

Independent Variables	Pre-Enrollment Model	Census Date Model	End of First Semester Model
Family Background/Individual Characteristics			
Age	0.89	0.91	0.86*
Asian?	1.63**	1.59**	1.72**
Female?	-----	-----	0.81
New York City Resident?	2.08**	1.66**	1.65**
Pre-College Schooling			
High School Grade Point Average	1.61**	1.41˙	-----
Number of Transfer/AP Credits Accepted	-----	1.05*	1.05**
Institutional Commitment			
Attended Summer Orientation?	1.61**	1.56**	1.35*
Academic Integration (all as of first term)			
Cumulative Attempted Hours	-----	1.14**	1.15**
Unearned Points	-----	-----	0.90**
Semester Grade Point Average	-----	-----	1.41*
Student's Major is Undecided?	0.75*	0.77˙	0.77˙
Social Integration Into College (first term)			
Did Student Meet with the Freshman Dean?	-----	-----	0.54**
Finances (first year)			
Amount of Institutional Grant Aid Awarded	-----	1.00*	1.00
Amount of Unmet Financial Need	1.00*	-----	1.00*
Bursar Balance at End of Third Week	-----	1.00*	1.00**
Receiving Tuition Remission Benefits?	2.95˙	2.67	-----
% of Need Met by Financial Aid Package	-----	1.29	-----
Interactions			
Unmet Financial Need * New York City Resident	1.00	-----	-----
Semester GPA * Unearned Points	-----	-----	0.98
# of Transfer/AP Points * Amount of Grant Aid	-----	1.00	1.00
Amount of Grant Aid * % of Need Met	-----	1.00*	-----
Goodness of Fit	2161.35	2168.77	2222.97
-2 Log Likelihood	1558.95	1524.54	1419.47
χ^2	79.54	113.72	228.97
Significance of χ^2	p≤.0001	p≤.0001	p≤.0001

Sample size for the pre-enrollment and census date models is 2201. Sample size for the end of first semester model is 2209. **p≤.01, *p≤.05, ˙p≤.10.

BEST COPY AVAILABLE

did not increase or decrease the overall odds that the student would be retained into the second year.

The final model represents variables that are available after the end of the first semester. In this model, we found that three positive predictors dropped out of the logit equation: high school grade point average; the percentage of need met by the financial aid package; and whether the student was receiving tuition remission benefits. Four variables that were not previously present in the two earlier models were added, and all but one decreased the overall odds that the student would be retained into the second year. The negative influences on retention are: being female; having a higher number of unearned hours; and meeting with the Freshman Dean. This last variable, meeting with the Freshman Dean, requires further explanation. Generally, students are "invited" to see the Freshman Dean if they are having difficulty or appear to be at risk. These meetings may come about as the result of a faculty member suggesting to the Freshman Dean that a student is having difficulty, from review of mid-term grades, or in other, informal ways. Based on this, it is not surprising that "meeting with the Freshman Dean" is negatively associated with retention.

On the other hand, having a higher first semester grade point average increased the odds that the student would be retained into the second year. Five variables contributed to the overall fit of the logit model, but did not really increase or decrease the odds that the student would be retained into the second year. They are the amount of institutional grant aid awarded; the amount of unmet financial need; the student's bursar balance at the end of the third week; as well as the interaction between semester grade point average and the number of unearned points and the interaction between the number of transfer/advanced placement credit hours and the amount of institutional grant aid received.

The goal of this project was to classify students as projected retainers or projected attriters to the second fall. Classification is possible with logistic regression because the ultimate result of the regression equation is a probability, in this case, the probability that a student will be retained. The probability can range from zero to 1, with the most typical classification scheme being where observations with estimated probabilities less than 0.5 are classified as not occurring while those observations with estimated probabilities of 0.5 and greater are classified as occurring.

There is a large disparity between the number of students in the attrited group (272) versus the number in the retained group (1,937) in this study. Table 3 shows that by using the standard classification table, where the most common "cut point" of 0.5 is used, virtually all of the students who retained were correctly classified (99% to 100%), while only 0.4% to 14.3% (depending on the model used) of those who attrited were correctly classified. This is due to the fact that in logistic regression, "classification is sensitive to the relative sizes of the two component groups and will always favor classification into the larger group." (Hosmer and Lemeshow, 1989, p. 147)

To overcome this problem, we explored using probability "cut points" ranging from 0.5 to 0.85 to determine classification for the models. In other words, rather than saying those students with estimated probabilities of less than 0.5 are categorized as projected attriters while those with estimated probabilities of 0.5 or greater are categorized as projected retainers, one can set the decision point at some other value, for example 0.7. In this case, students with estimated probabilities of less than 0.7 would be classified as projected attriters while students with estimated probabilities of 0.7 or greater would be classified as projected retainers. To find the optimal cut point, we analyzed the result of using a number of different values. Some of the classifications produced by these various cut points significantly improve the percentage of

Table 3

Logistic Regression Results--Classification Results

% Categorized Correctly at Various Probability "Cut Points"	Pre-Enrollment Model	Census Date Model	End of First Semester Model
Cut Point of 0.50			
% of attrited correctly predicted	0.4%	0.7%	14.3%
% of retained correctly predicted	100.0%	100.0%	99.0%
Concordant Predictions	87.8%	87.8%	88.6%
Cut Point of 0.60			
% of attrited correctly predicted	0.7%	1.9%	21.0%
% of retained correctly predicted	100.0%	99.6%	98.3%
Concordant Predictions	87.8%	87.6%	88.8%
Cut Point of 0.70			
% of attrited correctly predicted	3.3%	7.4%	27.9%
% of retained correctly predicted	99.2%	98.0%	96.4%
Concordant Predictions	87.4%	86.9%	88.0%
Cut Point of 0.80			
% of attrited correctly predicted	18.9%	30.4%	39.3%
% of retained correctly predicted	90.6%	88.2%	90.9%
Concordant Predictions	81.8%	81.1%	84.6%
Cut Point of 0.85			
% of attrited correctly predicted	47.0%	54.8%	51.8%
% of retained correctly predicted	73.0%	72.2%	82.0%
Concordant Predictions	70.0%	70.1%	78.3%
Goodness of Fit	2161.35	2168.77	2222.97
-2LL Likelihood	1558.95	1524.54	1419.47
χ^2	79.54	113.72	228.97
Significance of χ^2	$p \leq .0001$	$p \leq .0001$	$p \leq .0001$

Sample size for the pre-enrollment and census date models is 2201. Sample size for the end of first semester model is 2209.

attrited students correctly predicted, while not greatly reducing the percentage of retainers correctly predicted.

Implications

It is often said that data analysis is more of an art than a science, and this study is no exception. In this case we are looking at balancing many needs:

1. How early can we identify students who are risk? Can we identify them before they arrive on campus for their first semester? After the third week, when we know their academic program? At the end of the first semester, when the results from their first term of undergraduate work are known? The models become stronger the longer one waits, but this needs to be balanced against the fact that the earlier students can be identified and an intervention organized, the better the chance of having a successful intervention.
2. How accurately can we predict which students are at risk and which are not? As we adjust the "cut point" for mapping the estimated probability of retention to an attrition/retention prediction, we may improve on identifying the students who are at risk, at the cost of accurate predictions for those who will retain. The trade-off in this situation is being able to identify a reasonable portion of those students at risk, so that they can receive the intervention, versus the cost of providing the intervention to those who are actually not at risk, due to misclassification.

Earlier in the paper we indicated that the goal of this project is to identify students who are at risk for leaving the University. Logically, there are many different points where this identification could take place. The most accurate classification would occur after the fact, when we know with certainty who has and who has not registered for the second Fall. Although

perfectly accurate, this probably would not help the college with its early intervention program. As we move earlier in time, the accuracy of the predictions will decrease, but the ability to intervene in a timely fashion increases.

But instead of suggesting that there is an optimal answer as to which model to adopt (pre-enrollment, census date or end of first semester), we would suggest that the best approach would be to use all three models to plan an intervention strategy. In other words, before the new students arrive on campus, the pre-enrollment model can be run to identify students who are at risk. The school can then opt to intervene with these students in a manner appropriate for that time frame, perhaps sending personal invitations from the Dean to an informal reception with faculty members, or perhaps simply a listing of the support services which are available on campus.

Then, once census date came and enrollment information was available, the second model could be run to identify students at risk. It should be noted that there will not be perfect overlap between the two sets of students identified as being at risk. Some students who were identified as being at risk from the pre-enrollment model will not be classified that way based on the second model, while there will be other students who were not initially classified as being at risk, but who, based on enrollment information, are at-risk. We would see the optimal strategy as one that would take the union of the two sets of students identified and treat them together as a set of students at risk. Interventions at this point might include invitations to freshman programming events or identification for advisors of students whose progress should be tracked more closely, perhaps by a phone call from the advisor to the student to see how things are going or to offer assistance, as well as contact between the advisor and the students' instructors, to monitor progress.

Similarly, once the first semester is complete, the end of semester model can be run, which will identify a third set of students who are at risk. Again, there will be some overlap among the three groups of students, and one approach would simply be to treat the union of all the groups as the students at risk. Schools can determine what would be the most appropriate intervention at this point, which might range from blocking registration until the student sees an advisor to phone calls to written communications.

Regardless of the particular approach a college prefers, both in terms of when to run attrition models as well as in terms of the type of intervention, we now have a means to identify those students who should be targeted. We have also indicated that, depending on the circumstances at individual campuses, the decision of which model to use (pre-enrollment, census data, or end of first semester), or whether to use two or three, is an individual, campus specific choice. But we have one more area to review, that of deciding where to establish the cut-points for the probabilities.

One measure of how well the model does in classifying students is simply the percent of predictions which a model accurately provides, when applied to the data used to construct the model. This measure is known as concordant predictions. Often, a goal in logistic regression is to try and have the highest number of concordant predictions. If we look at Table 3 for the Census Date model, we see the largest number of correct predictions, 87.8%, occurs using a cut point of 0.5. In this case, this high concordancy is driven by the fact that 100% of the students being retained are accurately predicted, while only 0.7% of the attriters have been correctly identified. Although overall the model works well, for our purpose, identifying students at risk, the model with the 0.5 cut point has little value.

Therefore, we need to look beyond the percent of concordant predictions in determining what cut point to use. Still using the Census Date model, we see that as the cut point

increases, so does the percent of attriters who are correctly identified. Unfortunately, even with the cut point set at 0.85, we still only identify 54.8% of the attriters; the rest would not receive the desired intervention. But, 54.8% (over half) is a much better level of identification than at the default cut point of 0.5, where only 0.7% of the attriters were found.

We see, however, that the concordant predictions have dropped to 70.1%. A large part of that drop is due to the decrease in the accuracy of predicting those who retain. In practical terms, this means that students who retain to the second Fall would be identified as at risk, and would receive the intervention aimed at those most likely to attrit.

Is this a problem? We would argue not, for a number of reasons. First, the cost of many interventions is low, and so exposing students who are not at risk to them will not likely create a large expense. Further, we need to remember that we are looking at a very narrow view of attrition: those students who are out of attendance in their second fall. It is not unreasonable to hypothesize that some of the students, identified as at risk but actually in attendance in the second fall, leave the University sometime after that. If that is correct, then exposing such students to any intervention geared to retaining to students may produce longer term dividends.

The point here really is that a simple measure of the overall accuracy of the predictions is not sufficient for our purposes. We need to accept the possibility that we will not have the highest level of accurate predictions, as a trade off for identifying more students who are at risk.

Further Research

We recognize that we, ourselves, are just scratching the surface in analyzing student attrition and retention patterns. We look forward to expanding our research to the other undergraduate (and eventually graduate) schools of the University. Along the way, we would also like to enhance the model we've begun with here, by examining such data as high school

size; hometown size; high school curriculum; distance to New York City from the student's home; whether or not the student has an on-campus job; and whether or not the student was on a waiting list for courses, to see if any of these additional variables are statistically relevant to predicting whether or not a student will remain at NYU.

Additionally, we recognize that our analysis may be confounded by there being two distinct types of students who leave NYU: high ability students who have used NYU to enhance their preparation and who might transfer to other high caliber institutions, and low ability students who either transfer to a less rigorous institution or who abandon higher education altogether. The attributes of these two groups are most likely different on a variety of variables (level of SAT score; high school GPA; first term college GPA), and so relationships between these variables and retention may not be clear. We would like to find out more about the students who leave, a notoriously difficult exercise, to see if knowing what a student does after leaving NYU can enhance the prediction of students at risk. We could also imagine where knowing why a student was at risk (high ability versus low ability) would lead to different types of interventions, for example invitations to conduct research with faculty members versus invitations to a study skills workshop.

Summary

We have presented the approach taken at New York University to better understand who are the students at risk for not re-enrolling in their second fall semester. The variables presented are institution specific; we used, and presented here, the data we had available, recognizing that other institutions may have other data available, at various times during the semester. We have developed three models, which can be used at different times, and suggest that it might make sense to use all three to identify students at risk. Finally, we have

explained how to hone the model, perhaps not to produce the most "correct" predictions, but to be the most useful model for our purposes.

References

Cabrera, A., Nora, A., & Casañeda, M. (1992). The role of finances in the persistence process: a structural model. Research in Higher Education, 33, 571-594.

Hosmer, D. and Lemeshow, S. (1989). Applied Logistic Regression. New York: John Wiley and Sons.

Pascarella, E. and Terenzini, P. (1991). How College Affects Students. San Francisco: Jossey Bass.

Spady, W. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. Interchange, 1, 64-85.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45, 89-125.